

# A Critical Examination of RESCAL for Completion of Knowledge Bases with Transitive Relations

Pushpendre Rastogi<sup>†</sup> and Benjamin Van Durme  
Johns Hopkins University

## Abstract

Link prediction in large knowledge graphs has received a lot of attention recently because of its importance for inferring missing relations and for completing and improving noisily extracted knowledge graphs. Over the years a number of machine learning researchers have presented various models for predicting the presence of missing relations in a knowledge base. Although all the previous methods are presented with empirical results that show high performance on select datasets, there is almost no previous work on understanding the connection between properties of a knowledge base and the performance of a model. In this paper we analyze the RESCAL method (Nickel et al., 2011) and show that it can not encode *asymmetric transitive* relations in knowledge bases.

## 1 Introduction

Large-scale and highly accurate knowledge bases (KB) such as Freebase (Bollacker et al., 2008) and YAGO2 (Hoffart et al., 2013), have come to be recognized as essential for high performance on various Natural Language Processing (NLP) tasks. Relation extraction, Question Answering (Dalton et al., 2014; Fader et al., 2014; Yao and Van Durme, 2014) and Entity Recognition/Disambiguation in informal domains (Ritter et al., 2011; Zheng et al., 2012) are a few examples of tasks where KBs have proved to be invaluable. As these examples demonstrate, increasing the recall of knowledge bases without compromising on the precision has a direct impact on several tasks that are the focus of NLP research. Because of this importance of high recall in knowledge bases

and because the recall of even Freebase, the largest open source KB, is still quite low<sup>1</sup> a number of researchers have published heuristics with their empirical performance on automatically inferring the information that is missing in knowledge bases. Unfortunately the literature on theoretical analysis for these methods is still scarce.

In this paper we analyze RESCAL (Nickel et al., 2011) which is a widely cited method for inferring missing relations in KBs. The RESCAL method embeds entities and relations in a KB using vectors and matrices respectively and it predicts the true status of an edge between two nodes using these representations. Although RESCAL was introduced in 2011 and has been shown to be effective on a variety of datasets (Toutanova et al., 2015; Nickel et al., 2011; Nickel et al., 2012) there has been no theoretical analysis of the failure modes of this method. We show, both theoretically and experimentally (Sections 2 and 3), that RESCAL is not suitable for predicting missing relations in a KB that contains transitive and asymmetric relations such as the “type of” relation which is very important in Freebase (Guha, 2015) and the “hypernym” relation which is important in WordNet (Miller, 1995).

## 2 Analysis of RESCAL

**Notation:** A knowledge base contains, but is not equal to, a collection of (*subject, relation, object*) triples. Each triple encodes the fact that a *subject* entity is related to an *object* through a particular type of *relation*. Let  $\mathcal{V}$  and  $\mathcal{R}$  denote the finite set of entities and relationships. We assume that  $\mathcal{R}$  includes a type for the *null relation* or *no relation*. Let  $V = |\mathcal{V}|$  and  $R = |\mathcal{R}|$  denote the number of entities and relations. We use  $v$  and  $r$  to denote a

<sup>1</sup>It was reported by Dong et al. (2014) in October 2013, that 71% of people in Freebase had no known place of birth and that 75% had no known nationality.

<sup>†</sup>pushpendre@jhu.edu

generic entity and relation respectively. The shorthand  $[n]$  denotes  $\{x | 1 \leq x \leq n, x \in \mathbb{N}\}$ . Let  $E$  be the number of triples known to us and let  $e$  denote a generic triple. We denote the subject, object and relation of  $e$  through  $e^{sub} \in \mathcal{V}$ ,  $e^{obj} \in \mathcal{V}$  and  $e^{rel} \in \mathcal{R}$  respectively and we denote the entire collection of facts as  $\mathcal{E} = \{e_k | k \in [E]\}$ .

**RESCAL:** The RESCAL model associates each entity  $v$  with the vector  $a_v \in \mathbb{R}^d$  and it represents the relation  $r$  through the matrix  $M_r \in \mathbb{R}^{d \times d}$ . Let  $v$  and  $v'$  denote two entities whose relationship is unknown, the RESCAL model predicts the relation between  $v$  and  $v'$  to be:

$$\hat{r} = \underset{r \in \mathcal{R}}{\operatorname{argmax}} s(v, r, v') \quad (1)$$

$$s(v, r, v') = s(v, r, v') = a_v^T M_r a_{v'} \quad (2)$$

Note that in general if the matrix  $M_r$  is asymmetric then the score function  $s$  would also be asymmetric, i.e.,  $s(v, r, v') \neq s(v', r, v)$ . Let  $\Theta = \{a_v | v \in \mathcal{V}\} \cup \{M_r | r \in \mathcal{R}\}$ . Clearly  $\Theta$  parameterizes RESCAL. Therefore even though the same embedding is used to represent an entity regardless of whether it is the first or the second entity in a relation, the RESCAL model could still handle asymmetric relations if the matrix  $M_r$  is asymmetric.

**Transitive Relations and RESCAL:** In addition to relational information about the binary connections between entities, many KBs contain information about the relations themselves. For example, consider the toy knowledge base depicted in Figure 1. Based on the information that *Fluffy is-a Dog* and that a *Dog is-a Animal* and that *is-a* is a transitive relations we can infer missing relations such as *Fluffy is-a Animal*.

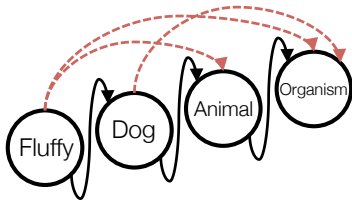


Figure 1: A toy knowledge base containing only *is-a* relations. The dashed edges indicate unobserved relations that can be recovered using the observed edges and the fact that *is-a* is a transitive relation.

Let us now analyze what happens when we encode a transitive, asymmetric relation with RESCAL. Consider the situation where the set  $\mathcal{R}$

only contains two relations  $\{r_0, r_1\}$ .  $r_1$  denotes the presence of the *is-a* relation and  $r_0$  denotes the absence of that relation. The RESCAL model can only follow the chain of transitive relations and infer missing edges using existing information in the graph if for all triples of vertices  $v, v', v''$  in  $\mathcal{V}$  for which we have observed  $(v, \text{is-a}, v')$  and  $(v', \text{is-a}, v'')$  the following holds true:

$$s(v, r_1, v') > s(v, r_0, v') \wedge s(v', r_1, v'') > s(v', r_0, v'') \\ \implies s(v, r_1, v'') > s(v, r_0, v'')$$

This can be rewritten as:

$$\forall v, v', v'' \in \mathcal{V}$$

$$a_v^T (M_{r_1} - M_{r_0}) a_{v'} > 0 \wedge a_{v'}^T (M_{r_1} - M_{r_0}) a_{v''} > 0 \\ \implies a_v^T (M_{r_1} - M_{r_0}) a_{v''} > 0 \quad (3)$$

We now define a *transitive matrix* and state a theorem that we prove in the Appendix.

**Definition** We say that a matrix  $M \in \mathbb{R}^{d \times d}$  is transitive if every triple of vectors  $a, b, c \in \mathbb{R}^d$  that satisfy  $a^T M b > 0$  and  $b^T M c > 0$  also satisfy  $a^T M c > 0$ .

**Theorem 1.** *Every transitive matrix is symmetric.*

If we enforce the constraint in Equation 3 to hold for all possible vectors and not just a finite number of vectors then  $M_{r_1} - M_{r_0}$  is a transitive matrix. By Theorem 1  $M_{r_1} - M_{r_0}$  must be symmetric. This implies that if the RESCAL model predicts that  $s(v, r_1, v') > s(v, r_0, v')$  then it would also predict that  $s(v', r_1, v) > s(v', r_0, v)$ . In terms of the toy KB shown in Figure 1; if the RESCAL model predicts that *Fluffy is-a Animal* then it would also predict that *Animal is-a Fluffy*. Therefore the RESCAL model is not suitable for encoding asymmetric, transitive relations.

### 3 Experiments

During our analysis in Section 2, we made assumption that the constraint of equation 3 held over all vectors in  $\mathbb{R}^d$  instead of just a finite number of vector triples. This assumption was used to make conclusions about RESCAL using Theorem 1.

A fair criticism of our analysis is that practically the RESCAL model only needs to encode a finite number of vertices into vector space and it is possible that there exists an asymmetric matrix that can correctly make the finite number of deductions that are possible inside a finite KB. This could be

especially true when the dimensionality  $d$  of the RESCAL embeddings is high. On the other hand, it is intuitive that as the number of entities inside a KB increases our assumptions and analysis would become increasingly better approximations of reality. Therefore the performance of the RESCAL model should degrade as the number of entities inside the KB increases and the dimensionality of the embeddings remains constant.

### 3.1 On Simulated Data

In order to test the applicability of our analysis we performed the following experiment: We started with a complete, balanced, rooted, directed binary tree  $\mathcal{T}$ , with edges directed *from* the root *to* its children. We then augmented  $\mathcal{T}$  as follows: For every tuple of distinct vertices  $v, v'$  we added a new edge to  $\mathcal{T}$  if there already existed a directed path starting at  $v$  and ending at  $v'$  in  $\mathcal{T}$ . We stopped when we could not add any more edges without creating multi-edges. For the rest of the paper we denote this resulting set of ordered pairs of vertices as  $\mathcal{E}$  and those pairs of vertices that are not in  $\mathcal{E}$  as  $\mathcal{E}^c$ . For example  $\mathcal{E}$  contains an edge from the root vertex to every other vertex and  $\mathcal{E}^c$  contains an edge from every vertex to the root vertex. For a tree of depth 11,  $V = 2047$ ,  $E = 18,434$  and  $|\mathcal{E}^c| = 4,171,775$ .

We trained the RESCAL model under two settings: In the first setting we used entire  $\mathcal{E}$  and  $\mathcal{E}^c$  as training inputs to the RESCAL model. We denote this setup as *FullSet*. In the second setting we randomly sample  $\mathcal{E}^c$  and select only  $E = |\mathcal{E}|$  edges from  $\mathcal{E}^c$ . We denote this training setup as *SubSet*. Note however, that all in the edges in  $\mathcal{E}$  including all the edges in the original tree are always used during both *FullSet* and *SubSet*.

For both the settings of *FullSet* and *SubSet* we trained the RESCAL model 5 times and evaluated the models' predictions on the following three subsets of the edges:  $\mathcal{E}$ ,  $\mathcal{E}^c$  and  $\mathcal{E}^{(rev)}$ .  $\mathcal{E}$  and  $\mathcal{E}^c$  were introduced earlier. To recall,  $\mathcal{E}$  contains all ordered pairs of vertices that are in the transitive relation of being connected,  $\mathcal{E}^c$  contains pairs of vertices that are not connected and not in a relation.  $\mathcal{E}^{(rev)}$  denotes the set of ordered pairs whose reverse pair exists in  $\mathcal{E}$ . I.e.,  $\mathcal{E}^{rev} = \{(u, v) | (v, u) \in \mathcal{E}\}$ . For every edge in each of these subsets we evaluate the model's performance under 0–1 loss. For example, when we evaluate the performance of RESCAL on an edge  $(v, v') \in \mathcal{E}$  we evaluate whether the model

assigns a higher score to  $(v, r_1, v')$  than  $(v, r_0, v')$  and reward the model by 1 point if it makes the right prediction and 0 otherwise. As before,  $r_1$  and  $r_0$  denote the presence and absence of relationship between  $v$  and  $v'$ .

We note that low Performance on  $\mathcal{E}^{rev}$  and high performance on  $\mathcal{E}$  would indicate exactly the kind of failure that we predicted from our analysis.

As explained earlier, the dimensionality of the RESCAL embedding,  $d$ , and the number of entities,  $V$  significantly influence the performance of RESCAL therefore we vary them and tabulate the results in Table 1 and 2. [upload\\_to\\_arxiv](#)

$d$	$V = 2047$	4095	8191
50	66 100 100	60 100 100	54 100 100
100	76 100 100	69 100 100	63 100 100
200	86 100 100	79 100 100	72 100 100
400	94 100 100	88 100 100	81 100 100

Table 1: Percentage accuracy of RESCAL with *FullSet*. Every table element is a triple of numbers measuring the performance of RESCAL on  $\mathcal{E}$ ,  $\mathcal{E}^c$ ,  $\mathcal{E}^{rev}$  respectively.  $V$  denotes the number of nodes in the tree and  $d$  denotes the number of dimensions used to parameterize the entities.

$d$	$V = 2047$	4095	8191
50	100 93 52	100 91 48	100 89 44
100	100 78 58	100 92 56	100 89 52
200	100 60 72	100 71 61	100 90 59
400	100 54 67	100 57 70	100 65 62

Table 2: Accuracy of RESCAL trained with *SubSet*.

### 3.2 On WordNet

In order to test our analysis on real data we performed experiments on the WordNet dataset. WordNet contains vertices called synsets that are arranged in a tree like hierarchy under the relation of *hyponymy*. For example, a *dog* is a hyponym of *animal* and an animal is a hyponym of *living-thing* therefore a *dog* is a hyponym of *living-thing*. To conduct our experiments we extracted all the hyponyms of the *living-thing* synset as a tree and edges to that tree to form a transitive closure under the hyponym relation. The *living-thing* synset contained 16255 hyponyms which were connected with 16489 edges and after performing the transitive closure the number of edges became 128241, i.e.,  $V = 16,255$  and  $E = 128,241$ . We performed two experiments under the *FullSet* and *SubSet* protocols in exactly the same way as described

in Section 3.1 with the new graph. The results, shown in Table 3, exhibit the same trends as seen in Table 1 and 2. See the following section for a more thorough discussion of results.

$d$	<i>FullSet</i>	<i>SubSet</i>
50	71 100 100	100 93 58
100	79 100 100	100 94 60
200	84 100 100	100 93 63
400	89 100 100	100 68 69

Table 3: Results from experiments on WordNet. Specifically we chose to use the subtree rooted at the *living\_things* synset from the WordNet hierarchy. Every synset in the subtree corresponds to a vertex. Consequently, for all our experiments  $V = 16413$ .

## 4 Related Work

Most previous works for inferring the missing information in knowledge bases assumes that a knowledge base is just a graph with labeled vertices and labeled edges (Nickel et al., 2016; Toutanova et al., 2015) and they either focus on inferring which labeled edge, if any, should be used to connect two previously unconnected vertices or they try to learn what vertex label/entity type, if any, should be used to annotate an unlabeled entity.

The task of predicting missing edges in a KB, which we focus on, has previously been called Link Prediction (Liben-Nowell and Kleinberg, 2007; Nickel et al., 2011), Knowledge Base Completion (KBC) (Socher et al., 2013; West et al., 2014) or more broadly Relational Machine Learning (Nickel et al., 2016).<sup>2</sup> Besides the before-mentioned papers the following publications also present models for KBC which we list without comments (Bordes et al., 2011; Lao et al., 2011; Gardner and Mitchell, 2015; Lin et al., 2015; Zhao et al., 2015; Wang et al., 2015; He et al., 2015; Wei et al., 2015).

## 5 Results and Discussion

Table 1 and 2 show the performance of the RESCAL model, for encoding three subsets of relational information,  $\mathcal{E}$ ,  $\mathcal{E}^c$  and  $\mathcal{E}^{rev}$  in increasingly

<sup>2</sup>The terminology used for the task of inferring missing vertex labels, which is not the focus of this paper, is even more diverse. This task has been termed Class/Labeled Instance acquisition (Van Durme and Pasca, 2008; Talukdar and Pereira, 2010), Collective Classification (Sen et al., 2008) and Vertex Nomination (Fishkind et al., 2015).

large KBs with a single transitive relation under a broad range of settings.

The results in Table 1 were obtained by feeding RESCAL  $\mathcal{E} \cup \mathcal{E}^c$  as training data. Note that RESCAL received all possible information during training so we are evaluating the training accuracy of the model at this point. Low accuracy under this setting implies that the model does not have the capacity to learn the rules in the knowledge base. We observe that the accuracy of RESCAL decreases as the number of entities,  $V$  increases and it increases as the dimensionality,  $d$  increases which is in line with our predictions. We also note that since  $\mathcal{E}^c$  is much larger than  $\mathcal{E}$  therefore the training objective of RESCAL favors good performance on  $\mathcal{E}^c$  and accordingly the accuracy of RESCAL on edges in  $\mathcal{E}^c$  remains high but the performance on  $\mathcal{E}$  suffers. The high accuracy of RESCAL with  $V = 2047$  and  $d = 400$  suggests that with a high enough dimensionality of the embeddings it is possible to embed a finite database with high accuracy. But increasing the dimensionality of RESCAL embeddings can become infeasible for an extremely large knowledge base. Also we can observe that the performance of RESCAL degrades as the number of entities inside the KB increases and the dimensionality of the embeddings remains constant.

The results in Table 2 were obtained by training RESCAL with  $\mathcal{E}$  and a subset of  $\mathcal{E}^c$ . This training method is closer to the way such embedding based methods for KBC are usually trained (Nickel et al., 2016). We observe that the accuracy of the RESCAL model on  $\mathcal{E}^{rev}$  is substantially lower than its performance on either  $\mathcal{E}$  or  $\mathcal{E}^c$ , especially in the upper triangle region of the table where  $V$  is high and  $d$  is low. This result is in accordance with our analysis that under the RESCAL mode if  $s(v, r_1, v') > s(v, r_0, v')$  then  $s(v', r_1, v) > s(v', r_0, v)$  as well. Our results also highlight a problem with the commonly employed KBC evaluation protocol of randomly dividing the edge set of a graph into train and test sets for measuring knowledge base completion accuracy. For example with  $d = 50$  the average accuracy on both  $\mathcal{E}$  and  $\mathcal{E}^c$  is quite high but on  $\mathcal{E}^{rev}$  accuracy is low even though  $\mathcal{E}^{rev}$  is a subset of  $\mathcal{E}^c$ . Such a failure would stay undetected with existing evaluation methods.

## 6 Conclusions

In this paper we investigated a popular KBC algorithm named RESCAL and through our analysis

of the scoring function employed in RESCAL and our experiments on simulated data, we showed that the RESCAL method does not perform well in encoding transitive and asymmetric relations and specifically that its inferences about edges that are the reverse of edges that are present in a knowledge base have a high chance of being incorrect. Although our analysis relied on somewhat strong assumptions that the constraint in equation 2 holds true over all points in the vector space we showed that the insights gained were useful in practice.

One of the key idea underlying our work was that knowledge bases should be considered as more than just graphs since KBs also contain logical structure amongst the predicates. By taking such logical structure, e.g., the constraint that if vertex  $v$  connects to  $v'$  and  $v'$  connects to  $v''$  then  $v$  connects to  $v''$ , to a logical extreme we came up with a well founded argument about the performance of RESCAL in encoding knowledge bases with transitive relations. We believe that this idea can be gainfully used to analyze other KBC methods as well.

## A Proof of Theorem 1

We note that Theorem 1 was first proven by Grinberg (2015). Here we give an alternative proof.

**Lemma 2.** *Every transitive matrix is PSD.*

*Proof.* Consider the triplet of vectors  $c := x, b := Mc, a := Mb$ . Then  $a^T(Mb) = ||Mb||^2 \geq 0$  and  $b^T(Mc) = ||b||^2 \geq 0$  and  $a^T Mc = b^T Mb$ . Either  $b = 0$  or  $b \neq 0$  and  $Mb = 0$ , or both  $Mb \neq 0$  and  $b \neq 0$  which implies  $b^T Mb > 0$  (by transitivity). In all three cases  $b^T Mb \geq 0$ .  $\square$

**Lemma 3.** *Let  $M_1, M_2 \in \mathbb{R}^{d \times d} \setminus \{0\}$ . If  $\forall x, y : x^T M_1 y > 0 \implies x^T M_2 y > 0$  then  $M_1 = \lambda M_2$  for some  $\lambda > 0$ .*

We defer the proof of this technical lemma to the supplementary material submitted with the paper.

**Lemma 4.** *If  $\exists x, y x^T M y > 0$  but  $x^T M^T y < 0$  then  $M$  is not transitive.*

*Proof.* Let  $x, y$  be two vectors that satisfy  $x^T M y > 0$  and  $x^T M^T y < 0$ . Since  $x^T M^T y = y^T M x$  therefore  $y^T M(-x) > 0$ . If we assume  $M$  is transitive, then  $x^T M(-x) > 0$  by transitivity, but Lemma 2 shows such an  $x$  cannot exist.  $\square$

**Theorem 1.** *Every transitive matrix is symmetric.*

*Proof.* By Lemma 4  $x^T M y > 0 \implies x^T M^T y > 0$ . Using Lemma 3 we get  $M = \lambda M^T$  for some  $\lambda > 0$ . Clearly  $\lambda = 1$ .  $\square$

## References

- [Bollacker et al.2008] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM.
- [Bordes et al.2011] Antoine Bordes, Jason Weston, Roman Collobert, and Yoshua Bengio. 2011. Learning structured embeddings of knowledge bases. In *AAAI Conference on Artificial Intelligence*.
- [Dalton et al.2014] Jeffrey Dalton, Laura Dietz, and James Allan. 2014. Entity query feature expansion using knowledge base links. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 365–374. ACM.
- [Dong et al.2014] Xin Dong, Evgeniy Gabrilovich, Jeremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmman, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 601–610. ACM.
- [Fader et al.2014] Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2014. Open question answering over curated and extracted knowledge bases. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1156–1165, New York, NY, USA. ACM.
- [Fishkind et al.2015] Donniell E Fishkind, Vince Lyzinski, Henry Pao, Li Chen, Carey E Priebe, et al. 2015. Vertex nomination schemes for membership prediction. *The Annals of Applied Statistics*, 9(3):1510–1532.
- [Gardner and Mitchell2015] Matt Gardner and Tom Mitchell. 2015. Efficient and expressive knowledge base completion using subgraph feature extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1488–1498, Lisbon, Portugal, September. Association for Computational Linguistics.
- [Grinberg2015] Darij Grinberg. 2015. Existence and characterization of transitive matrices? MathOverflow. URL:<http://mathoverflow.net/q/212808> (version: 2015-08-01).
- [Guha2015] Ramanathan Guha. 2015. Towards a model theory for distributed representations. In *AAAI Spring Symposium Series*.
- [He et al.2015] Wenqiang He, Yansong Feng, Lei Zou, and Dongyan Zhao. 2015. Knowledge base completion using matrix factorization. In *Web Technologies and Applications*, pages 256–267. Springer.

- [Hoffart et al.2013] Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence*, 194:28–61.
- [Lao et al.2011] Ni Lao, Tom Mitchell, and William W. Cohen. 2011. Random walk inference and learning in a large scale knowledge base. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11*, pages 529–539, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Liben-Nowell and Kleinberg2007] David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031.
- [Lin et al.2015] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *The 29th AAAI Conference on Artificial Intelligence*, pages 2181–2187. AAAI.
- [Miller1995] George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.
- [Nickel et al.2011] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 809–816.
- [Nickel et al.2012] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2012. Factorizing yago: scalable machine learning for linked data. In *Proceedings of the 21st international conference on World Wide Web*, pages 271–280. ACM.
- [Nickel et al.2016] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2016. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, Jan.
- [Ritter et al.2011] Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics.
- [Sen et al.2008] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. 2008. Collective classification in network data. *AI magazine*, 29(3):93.
- [Socher et al.2013] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems*, pages 926–934.
- [Talukdar and Pereira2010] Partha Pratim Talukdar and Fernando Pereira. 2010. Experiments in graph-based semi-supervised learning methods for class-instance acquisition. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1473–1481. Association for Computational Linguistics.
- [Toutanova et al.2015] Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. 2015. Representing text for joint embedding of text and knowledge bases. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Lisbon, Portugal, September. Association for Computational Linguistics.
- [Van Durme and Pasca2008] Benjamin Van Durme and Marius Pasca. 2008. Finding cars, goddesses and enzymes: Parametrizable acquisition of labeled instances for open-domain information extraction. In *AAAI*, volume 8, pages 1243–1248.
- [Wang et al.2015] Quan Wang, Bin Wang, and Li Guo. 2015. Knowledge base completion using embeddings and rules. In *Proceedings of the 24th International Conference on Artificial Intelligence, IJCAI'15*, pages 1859–1865. AAAI Press.
- [Wei et al.2015] Zhuoyu Wei, Jun Zhao, Kang Liu, Zhenyu Qi, Zhengya Sun, and Guanhua Tian. 2015. Large-scale knowledge base completion: Inferring via grounding network sampling over selected instances. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15*, pages 1331–1340, New York, NY, USA. ACM.
- [West et al.2014] Robert West, Evgeniy Gabrilovich, Kevin Murphy, Shaohua Sun, Rahul Gupta, and Dekang Lin. 2014. Knowledge base completion via search-based question answering. In *WWW*.
- [Yao and Van Durme2014] Xuchen Yao and Benjamin Van Durme. 2014. Information extraction over structured data: Question answering with freebase. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 956–966, Baltimore, Maryland, June. Association for Computational Linguistics.
- [Zhao et al.2015] Yu Zhao, Sheng Gao, Patrick Gallinari, and Jun Guo. 2015. Knowledge base completion by learning pairwise-interaction differentiated embeddings. *Data Mining and Knowledge Discovery*, 29(5):1486–1504.
- [Zheng et al.2012] Zhicheng Zheng, Xiance Si, Fangtao Li, Edward Y Chang, and Xiaoyan Zhu. 2012. Entity disambiguation with freebase. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 82–89. IEEE Computer Society.

## Supplementary Material

Before proving Lemma 3 let us present its analogue for vectors.

**Lemma 5.** *Let  $x, y \in \mathbb{R}^d \setminus \{0\}$ . If  $\nexists z \in \mathbb{R}^d$  such that  $x^T z > 0$  and  $y^T z < 0$  then  $x = \lambda y$  for some  $\lambda > 0$ .*

*Proof.* If  $x = \lambda y$  then  $x^T y = \lambda y^T y$ . Since  $y^T y > 0$  therefore  $\lambda > 0$ . In the case that  $x \neq \lambda y$  then by Cauchy Schwartz inequality  $D := (x^T y)^2 - (x^T x)(y^T y) \neq 0$ . Consider the vector  $\alpha x + \beta y$  with  $\alpha = -\frac{x^T y + y^T y}{D}$  and  $\beta = \frac{x^T y + x^T x}{D}$ . It is easy to check that  $(\alpha x + \beta y)^T x$  and  $(\alpha x + \beta y)^T y$  equal 1 and  $-1$ , which contradicts the hypothesis.  $\square$

**Lemma 3.** *Let  $M_1, M_2 \in \mathbb{R}^{d \times d} \setminus \{0\}$ . If  $\forall x, y : x^T M_1 y > 0 \implies x^T M_2 y > 0$  then  $M_1 = \lambda M_2$  for some  $\lambda > 0$ .*

*Proof.* Choose an  $x \in \mathbb{R}^d$  for which  $x^T M_1 \neq 0$ . If such an  $x$  does not exist  $M_1 = 0$  in contradiction to the hypothesis. Note that if  $x^T M_1 y \neq 0$  then either  $x^T M_1 y$  or  $x^T M_1 - y$  would be positive. Since  $(x^T M_1)y > 0 \implies (x^T M_2)y > 0$  therefore  $\nexists y$  for which  $(x^T M_1)y > 0$  but  $(x^T M_2)y < 0$ . By Lemma 5  $x^T M_1 = \lambda_x x^T M_2$ . Furthermore from the proof of Lemma 5  $\lambda_x = \frac{x^T M_1 M_2 x}{x^T M_2 M_2 x}$  therefore  $\lambda_x$  is continuous with respect to  $x$ . Now we prove that  $\lambda_x$  is constant. Consider vectors  $x$  and  $\alpha x$ . As shown earlier,  $(\alpha x)^T M_1 = \lambda_{\alpha x} (\alpha x)^T M_2$ . But  $(\alpha x)^T M_1 = \alpha (x^T M_1) = \alpha \lambda_x x^T M_2$ . Therefore  $\lambda_{\alpha x} = \lambda_x$ . Since  $\lambda_x$  is continuous at 0 therefore  $\lambda_{\alpha x}$  equals the constant  $\lambda_0$ . This implies  $x^T (M_1 - \lambda_0 M_2) = 0$ . Clearly  $\lambda = \lambda_0 > 0$ .  $\square$